University of Michigan School of Public Health

The University of Michigan Department of Biostatistics Working Paper Series

Year 2015 *Paper* 118

A Pairwise Likelihood Augmented Estimator for the Cox Model Under Left-Truncation

Fan Wu* Sehee Kim† Jing Qin‡

Rajiv Saran** Yi Li††

*University of Michigan - Ann Arbor, fannwu@umich.edu

†University of Michigan - Ann Arbor, seheek@umich.edu

‡National Institute of Allergy and Infectious Diseases

**University of Michigan - Ann Arbor

††University of Michigan - Ann Arbor, yili@umich.edu

A Pairwise Likelihood Augmented Estimator for the Cox Model Under Left-Truncation

Fan Wu, Sehee Kim, Jing Qin, Rajiv Saran, and Yi Li

Abstract

Survival data collected from prevalent cohorts are subject to left-truncation and the analysis is challenging. Conditional approaches for left-truncated data under the Cox model are inefficient as they typically ignore the information in the marginal likelihood of the truncation times. Length-biased sampling methods can improve the estimation efficiency but only when the stationarity assumption of the disease incidence holds, i.e., the truncation distribution is uniform; otherwise they may generate biased estimates. In this paper, we propose a semi-parametric method for the Cox model under general left-truncation, where the truncation distribution is unspecified. Our approach is to make inference based on the conditional likelihood augmented with a pairwise likelihood which eliminates the unspecified truncation distribution, yet retains the information about the regression coefficients and the baseline hazard function in the marginal likelihood. An iterative algorithm is provided to solve for the regression coefficients and the baseline hazard simultaneously. The proposed estimator is consistent and asymptotically normal with a closed-form consistent variance estimator. Simulations show a substantial efficiency gain in both the regression coefficients and the cumulative baseline hazard over the conditional approach estimator. Even when the stationarity assumption holds, our estimator results in better efficiency than some length-biased sampling estimators. An application to the analysis of a chronic kidney disease cohort study illustrates the utility of the method.

A Pairwise Likelihood Augmented Estimator for the Cox Model Under Left-Truncation

Fan Wu
Department of Biostatistics,
University of Michigan, Ann Arbor
and
Sehee Kim
Department of Biostatistics,
University of Michigan, Ann Arbor
and
Jing Qin
Biostatistics Research Branch,
National Institute of Allergy and Infectious Diseases
and
Rajiv Saran
Kidney Epidemiology and Cost Center,
University of Michigan, Ann Arbor
and
Yi Li *
Department of Biostatistics,
University of Michigan, Ann Arbor

March 24, 2016

*E-mail: yili@umich.edu

Abstract

Survival data collected from prevalent cohorts are subject to left-truncation and the analysis is challenging. Conditional approaches for left-truncated data under the Cox model are inefficient as they typically ignore the information in the marginal likelihood of the truncation times. Length-biased sampling methods can improve the estimation efficiency but only when the stationarity assumption of the disease incidence holds, i.e., the truncation distribution is uniform; otherwise they may generate biased estimates. In this paper, we propose a semiparametric method for the Cox model under general left-truncation, where the truncation distribution is unspecified. Our approach is to make inference based on the conditional likelihood augmented with a pairwise likelihood which eliminates the unspecified truncation distribution, yet retains the information about the regression coefficients and the baseline hazard function in the marginal likelihood. An iterative algorithm is provided to solve for the regression coefficients and the baseline hazard simultaneously. The proposed estimator is consistent and asymptotically normal with a closed-form consistent variance estimator. Simulations show a substantial efficiency gain in both the regression coefficients and the cumulative baseline hazard over the conditional approach estimator. Even when the stationarity assumption holds, our estimator results in better efficiency than some length-biased sampling estimators. An application to the analysis of a chronic kidney disease cohort study illustrates the utility of the method.

Keywords: Composite Likelihood, Empirical Process, U-Process, Self-Consistency, Chronic Kidney Diseases

1 Introduction

Survival data collected from a prevalent cohort, who already have the disease under study at enrollment, are subject to left-truncation. This is because those who died with the disease before enrollment would have no chance to be selected, while patients in the prevalent cohort, having survived until the time of enrollment, are healthier on average. To avoid overestimating the survival, conventional approaches make inferences conditional on truncation times (Turnbull, 1976; Wang et al., 1986; Tsai et al., 1987; Kalbfleisch and Lawless, 1991; Wang et al., 1993). However, they disregard the information about the regression coefficients in the marginal likelihood of the truncation times, and hence efficiency loss is expected (Huang et al., 2012).

If we assume the incidence rate of the disease is stationary over time, i.e., that the underlying truncation time follows a uniform distribution, left-truncation reduces to length-biased sampling (Vardi, 1982, 1989), since the probability of a patient being selected into the prevalent cohort is proportional to the length of his or her survival time. An extensive literature exists on regression methods with length-biased data (Wang, 1996; Shen et al., 2009; Qin and Shen, 2010; Ning et al., 2011, 2014; Qin et al., 2011; Huang et al., 2012; Huang and Qin, 2012). Incorporating the information in the marginal likelihood of the truncation times, these methods generally lead to considerable improvement of efficiency in estimation. Nevertheless, when the stationarity assumption is violated, length-biased sampling methods will yield inconsistent estimates (Huang and Qin, 2012).

The motivating study is a multi-center prevalent cohort study of patients with moderate to advanced chronic kidney disease (CKD) (Perlman et al., 2003), sponsored by the Renal Research Institute (RRI-CKD). Subjects with glomerular filtration rate (GFR) less than or equal to $50\,\mathrm{ml/min/1.73\,m^2}$ were invited to participate in the study from June 2000 to January 2006. In general, CKD patients are referred to nephrologists to receive special care and treatments following the diagnosis. The investigators were interested in whether the baseline patient characteristics at referral were related to the disease progression to end-stage renal disease (ESRD) or death. Since only patients surviving beyond the point of reaching moderate stage CKD could be enrolled, the RRI-CKD data were left-truncated. However, the target population that the investigators tried to infer on includes the patients who were diagnosed with CKD but had died or progressed before enrolment. As shown in Section 4, both the statistical test and the graphical assessment in the RRI-CKD data indicated deviation from the stationarity assumption, which prompted us to seek a method avoiding the biases when using length-biased sampling methods, yet improving efficiency under general left-truncation.

Recently, Huang and Qin (2013) proposed a more efficient estimator for the additive hazards model under general left-truncation. They used a pairwise likelihood (Liang and Qin, 2000) of the truncation times to eliminate the unspecified truncation distribution. The additive hazards model, however, is less commonly used than the Cox model, and its interpretation may be unfamiliar to practitioners. Moreover, the challenge of applying the approach by Huang and Qin (2013) to the Cox model lies in the complicated way that the pairwise likelihood still involves the cumulative baseline hazard function, causing serious theoretical and computational difficulties.

To improve the efficiency in estimation, following the idea of Huang and Qin (2013), we propose to augment the Cox partial likelihood with a pairwise likelihood constructed from the marginal likelihood of the truncation times. We have achieved several important improvements. First, we have designed an NPMLE-type inference procedure to estimate the cumulative baseline hazard function along with the regression coefficients. Second, we have provided an iterative algorithm that explores the self-consistency of the non-parametric estimator and guarantees a computationally efficient implementation. Finally, with the asymptotic results proven by empirical process and U-process theories, we provide a closed-form consistent sandwich variance estimator of the parameter estimates. Our simulation studies show that efficiency of both the regression coefficients and the cumulative baseline hazard function is improved by using the proposed method, especially the former. In summary, the proposed method enjoys a more precise survival estimation when the sample size is small to moderate, as in most pilot studies. It is somewhat surprising to note that, when the stationarity assumption holds, the efficiency gain for the proposed estimator of the regression coefficients is even greater than that for the composite partial-likelihood (CPL) estimator (Huang and Qin, 2012) derived under the parametric assumption of the truncation distribution, which makes the proposed method more appealing.

The rest of this manuscript is organized as follows. In Section 2, we introduce the proposed pairwise-likelihood augmented estimator and the algorithm for implementation. Asymptotic properties of the estimator are also provided (with proofs given in the Appendix). Simulation study results are shown in Section 3, where we compare the finite sample performance of the proposed method with the competing methods, including the conditional approach (Wang et al., 1993) and the CPL methods by Huang and Qin (2012), originally developed for length-biased sampling cases. Application to the RRI-CKD data is presented in Section 4. The manuscript concludes with a discussion of the proposed methods and suggestions for future work.

2 Methods

2.1 Notations

For a patient from the target population, let T^* be the underlying survival time, which measures the time from the disease onset (e.g., the referral in the RRI-CKD study) to the event. An independent truncation time, denoted as A^*, measures the time from the disease onset to the study enrollment. In a prevalent cohort, as shown in Figure 1, we observe the pairs (A^*, T^*) such that the events happen after the enrollment; that is, we only have realizations from $(A, T) \equiv (A^*, T^*) \mid A^* \leqslant T^*$. Note that the sampling scheme induces a positive correlation between A and T in the biased sample. The residual survival time, $V \equiv T - A$ is subject to potential censoring by C; thus, what we can observe are $X = \min(A + C, T)$ and $\Delta = I(T \leqslant A + C)$, where $I(\cdot)$ denotes the indicator function.

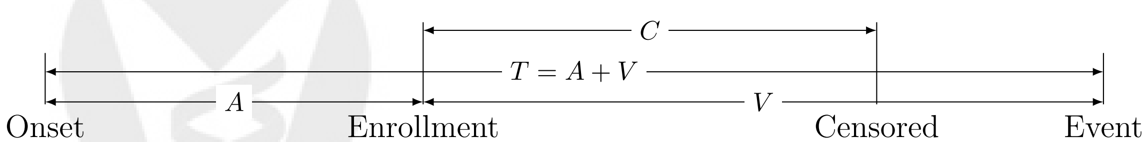

Figure 1: The structure of left-truncated data (adjusted from Shen et al., 2009).

We use f, S and λ to denote the density, survival and hazard functions of T^*, and the distribution function of A^* is denoted as G. Let $\mathbf{Z}$ be a $p \times 1$ vector of covariates for a subject in the prevalent cohort. A commonly used model that links the survival time T^*

to the covariates $\mathbf{Z}$ is the Cox proportional hazards model (Cox, 1972):

$$\lambda(t|\mathbf{Z};\boldsymbol{\beta}) = \lambda(t)\exp(\mathbf{Z}^T\boldsymbol{\beta}),$$

where $\lambda(\cdot)$ is an unspecified baseline hazard function, and $\boldsymbol{\beta}$ is a $p\times 1$ vector of regression coefficients. The baseline cumulative hazard function is defined as $\Lambda(\cdot) = \int_0^{\cdot} \lambda(s)ds$. Suppose we have independent and identically distributed (i.i.d.) observations $\mathcal{O}_i \equiv \{(A_i, X_i, \Delta_i, \mathbf{Z}_i), i = 1, \ldots, n\}$ on n individuals in the prevalent cohort study, under the assumption that C is independent of (A, T) given $\mathbf{Z}$, the full likelihood of the observed data is proportional to

$$\prod_{i=1}^{n} \Pr[A_i^*, T_i^*, C_i|\mathbf{Z}_i, A_i^* \leqslant T_i^*] \propto \prod_{i=1}^{n} \frac{f(X_i|\mathbf{Z}_i)^{\Delta_i} S(X_i|\mathbf{Z}_i)^{1-\Delta_i} dG(A_i)}{\int S(t|\mathbf{Z}_i)dG(t)} \equiv \mathcal{L}_n.$$

Note that we assume G does not depend on the covariates $\mathbf{Z}$. Unless otherwise specified, the integrals without the domain of integration are taken over the follow-up period $[0,\tau]$, where $0 < \tau < \infty$ is the maximum support of the observed survival time. The full likelihood can be further decomposed into two parts:

$$\mathcal{L}_n = \prod_{i=1}^{n} \frac{f(X_i|\mathbf{Z}_i)^{\Delta_i} S(X_i|\mathbf{Z}_i)^{1-\Delta_i}}{S(A_i|\mathbf{Z}_i)} \times \prod_{i=1}^{n} \frac{S(A_i|\mathbf{Z}_i)dG(A_i)}{\int S(t|\mathbf{Z}_i)dG(t)} \equiv \mathcal{L}_n^C \times \mathcal{L}_n^M, \tag{1}$$

where $\mathcal{L}_n^C$ is the conditional likelihood of (X, Δ) given $(A, \mathbf{Z})$, and $\mathcal{L}_n^M$ is the marginal likelihood of A given $\mathbf{Z}$ (Kalbfleisch and Sprott, 1970).

2.2 Pairwise-Likelihood Augmented Cox (PLAC) Estimator

In the presence of left-truncation, conditional inference based on $\mathcal{L}_n^C$ only, which uses the Cox's partial likelihood (Cox, 1975) with the modified at-risk indicator $Y_i(t) = I(A_i \leqslant t \leqslant X_i)$, has been proposed by Kalbfleisch and Lawless (1991) and Wang et al. (1993). From the classic results in Andersen and Gill (1982), the conditional approach yields consistent estimates, but it may be less efficient since it completely ignores the information about the parameters contained in $\mathcal{L}_n^M$.

When it is reasonable to assume the incidence of disease is stable over time (i.e., the stationary assumption), the probability of a subject being sampled is proportional to the length of his or her survival. Under this situation, the random truncation time is known to follow a uniform distribution. Taking advantage of the parametric form of G, length-biased sampling regression methods (Tsai, 2009; Qin and Shen, 2010; Huang and Qin, 2012) have been developed, which incorporate some information in $\mathcal{L}_n^M$ and hence result in estimators more efficient than that of the conditional approach. Among them, the composite partial-likelihood (CPL) method by Huang and Qin (2012) was shown to gain better or similar levels of efficiency compared with other competitors under the Cox model.

Deviating from the common length-biased sampling methods in the left-truncation literature, our method does not impose any parametric assumptions on the truncation time distribution function G, nor on the baseline hazard function λ. Our approach to improving efficiency is to supplement $\mathcal{L}_n^C$ with major information in $\mathcal{L}_n^M$ that depends on $\boldsymbol{\beta}$ and λ only, but to be free of G. Specifically, we first apply the pairwise pseudo-likelihood method by Liang and Qin (2000) to $\mathcal{L}_n^M$ in order to eliminate the nuisance parameter G, and then estimate $\boldsymbol{\beta}$ and λ based on a composite likelihood consisting of $\mathcal{L}_n^C$ and $\mathcal{L}_n^P$, where the pairwise pseudo-likelihood $\mathcal{L}_n^P$ is derived as follows.

Suppose a sample $\{(A_i, \mathbf{Z}_i), (A_j, \mathbf{Z}_j);\, i < j\}$ is available. Following the argument in Liang and Qin (2000), the pseudo-likelihood of the pair $(i,\, j)$, conditional on $(\mathbf{Z}_i, \mathbf{Z}_j)$ and the order statistic of (A_i, A_j), is given by

$$\frac{\frac{S(A_i|\mathbf{Z}_i)dG(A_i)}{\int S(t|\mathbf{Z}_i)dG(t)} \times \frac{S(A_j|\mathbf{Z}_j)dG(A_j)}{\int S(t|\mathbf{Z}_j)dG(t)}}{\frac{S(A_i|\mathbf{Z}_i)dG(A_i)}{\int S(t|\mathbf{Z}_i)dG(t)} \times \frac{S(A_j|\mathbf{Z}_j)dG(A_j)}{\int S(t|\mathbf{Z}_j)dG(t)} + \frac{S(A_i|\mathbf{Z}_j)dG(A_i)}{\int S(t|\mathbf{Z}_j)dG(t)} \times \frac{S(A_j|\mathbf{Z}_i)dG(A_j)}{\int S(t|\mathbf{Z}_i)dG(t)}} = \frac{1}{1 + R_{ij}(\boldsymbol{\beta}, \Lambda)},$$

where $R_{ij}(\boldsymbol{\beta}, \Lambda)$ denotes the generalized odds ratio and has the form

$$R_{ij}(\boldsymbol{\beta}, \Lambda) = \frac{S(A_i|\mathbf{Z}_j)S(A_j|\mathbf{Z}_i)}{S(A_i|\mathbf{Z}_i)S(A_j|\mathbf{Z}_j)} = \exp\left\{(e^{\mathbf{Z}_i^T\boldsymbol{\beta}} - e^{\mathbf{Z}_j^T\boldsymbol{\beta}})(\Lambda(A_i) - \Lambda(A_j))\right\} \quad (2)$$

under the Cox model. The pairwise likelihood $\mathcal{L}_n^P$ of all pairs is then given by

$$\mathcal{L}_n^P = \prod_{i<j}(1 + R_{ij}(\boldsymbol{\beta}, \Lambda))^{-1}.$$

It is worth noting that $\mathcal{L}_n^P$ is a function of $(\boldsymbol{\beta},\, \Lambda)$ only, not depending on G by canceling it out, whereas $\mathcal{L}_n^M$ is a function of $(\boldsymbol{\beta},\, \Lambda,\, G)$. An alternative approach would be to directly maximize the full likelihood $\mathcal{L}_n^C \times \mathcal{L}_n^M$ over $(\boldsymbol{\beta},\, \Lambda,\, G)$, which may be more efficient than the composite likelihood approach. However, when G is completely unspecified, maximizing over infinite dimensional parameters will increase computational cost and may not be stable numerically; thus, it is not worth pursuing when G is not the parameter of interest. Furthermore, simulation studies (Qin and Liang, 1999; Liang and Qin, 2000) show that the pairwise likelihood can retain the majority of the information in the likelihood from which it is derived, and that the efficiency loss may not be substantial, depending on the model as well as the values of the parameters. Therefore, to estimate $\boldsymbol{\beta}$ and λ, we propose using $\mathcal{L}_n^P$ as a reasonably good surrogate for $\mathcal{L}_n^M$ in the full likelihood approach. The analogous idea has been exploited in the additive hazards model by Huang and Qin (2013); however, the additive hazards model is less commonly used. Applying the pairwise-likelihood augmentation method to the Cox model will promote more practical use due to ease of interpretation to practitioners.

To account for the different magnitudes of $\log \mathcal{L}_n^C$ and $\log \mathcal{L}_n^P$ (there are n terms in $\log \mathcal{L}_n^C$ and $n(n-1)/2$ terms in $\log \mathcal{L}_n^P$), we maximize the following composite log-likelihood function:

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left\{\Delta_i\left(\log\lambda(X_i)+\mathbf{Z}_i^T\boldsymbol{\beta}\right)-\exp(\mathbf{Z}_i^T\boldsymbol{\beta})\int_0^\tau Y_i(t)\lambda(t)dt\right\}\\&-\frac{2}{n(n-1)}\sum_{i<j}\log\{1+R_{ij}(\boldsymbol{\beta},\Lambda)\},\end{aligned}$$

over the domain of $(\boldsymbol{\beta},\Lambda)$. Using the nonparametric maximum likelihood estimation (NPMLE) approach, we treat $\Lambda(\cdot)$ as a nondecreasing step function with jumps, denoted by $\Lambda\{\cdot\}$, only at the time points where events are observed and $\Lambda\,(0)=0$ (see Murphy et al., 1997; Zeng and Lin, 2006 among others). Let $w_1 < \cdots < w_m$ $(m \leqslant n)$ be the ordered distinct observed event times, and $\lambda_1 \equiv \Lambda\{w_1\}, \ldots\ \lambda_m \equiv \Lambda\{w_m\}$ be the corresponding positive jumps of Λ at these times. We denote by $\boldsymbol{\lambda} \equiv (\lambda_1, \ldots, \lambda_m)$ the vector of all positive jumps. For $k = 0,\, 1,\, 2$, we define the following functions which appear in $\log \mathcal{L}_n^P$ and its derivatives:

$$Q_{ij}^{(k)}(t;\boldsymbol{\beta})=\left(\mathbf{Z}_i^{\otimes k}e^{\mathbf{Z}_i^T\beta}-\mathbf{Z}_j^{\otimes k}e^{\mathbf{Z}_j^T\beta}\right)\left(I(t\leqslant A_i)-I(t\leqslant A_j)\right),$$

where $\mathbf{Z}^{\otimes 0} = 1$, $\mathbf{Z}^{\otimes 1} = \mathbf{Z}$, and $\mathbf{Z}^{\otimes 2} = \mathbf{Z}\mathbf{Z}^T$. Below we may suppress the dependence on model parameters, using R_{ij} and $Q_{ij}^{(k)}(t)$ to denote $R_{ij}(\boldsymbol{\beta}, \Lambda)$ and $Q_{ij}^{(k)}(t;\boldsymbol{\beta})$ when the meanings of the notations are clear from the context. Replacing $\lambda(t)$ with $\Lambda\{t\}$, we modify the composite log-likelihood as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$:

$$\begin{aligned} \ell_n^c(\boldsymbol{\beta}, \boldsymbol{\lambda}) &= \frac{1}{n}\sum_{i=1}^n \left\{ \Delta_i \left(\log \Lambda\{X_i\} + \mathbf{Z}_i^T \boldsymbol{\beta} \right) - \exp(\mathbf{Z}_i^T\boldsymbol{\beta}) \sum_{k=1}^m \lambda_k Y_i(w_k) \right\} \\ &\quad - \frac{2}{n(n-1)} \sum_{i<j} \log(1 + R_{ij}(\boldsymbol{\beta}, \boldsymbol{\lambda})), \end{aligned} \tag{3}$$

where

$$R_{ij}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \exp\left(\sum_{k=1}^m \lambda_k Q_{ij}^{(0)}(w_k) \right).$$

We refer to the resulting maximizer $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}})$ (or equivalently $(\hat{\boldsymbol{\beta}}, \hat{\Lambda})$) as the pairwise likelihood augmented Cox (PLAC) estimator, where Λ at a fixed time point $t \in [0, \tau]$ is estimated by $\hat{\Lambda}(t) = \sum_{k=1}^m \hat{\lambda}_k\, I(w_k \leqslant t)$. Specifically, differentiating (3) with respect to $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ yields the composite score functions (the dependence on n is suppressed):

$$\begin{aligned} U_{\beta}(\boldsymbol{\beta}, \boldsymbol{\lambda}) &= \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i \left\{ \Delta_i - e^{Z_i^T\beta} \sum_{k=1}^m \lambda_k Y_i(w_k) \right\} - \frac{1}{n(n-1)} \sum_{i \neq j} \frac{\sum_{k=1}^m \lambda_k Q_{ij}^{(1)}(w_k)}{1 + R_{ij}^{-1}}, \\ U_{\lambda_k}(\boldsymbol{\beta}, \boldsymbol{\lambda}) &= \frac{1}{n}\sum_{i=1}^n I(X_i = w_k) \left\{ \Delta_i/\lambda_k - Y_i(w_k)\, e^{\mathbf{Z}_i^T\beta} \right\} - \frac{1}{n(n-1)} \sum_{i \neq j} \frac{Q_{ij}^{(0)}(w_k)}{1 + R_{ij}^{-1}}. \end{aligned}$$

Let $U_{\boldsymbol{\lambda}}^T = (U_{\lambda_1}, \ldots, U_{\lambda_m})$, then the PLAC estimator $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}})$ is the solution to

$$U(\boldsymbol{\beta}, \boldsymbol{\lambda}) \equiv (U_{\boldsymbol{\beta}}^T, U_{\boldsymbol{\lambda}}^T)^T(\boldsymbol{\beta}, \boldsymbol{\lambda}) = 0, \tag{4}$$

which can be obtained numerically using the following algorithm, for example.

Unlike the conditional approach, directly solving the nonlinear system (4) is a difficult problem due to the computational complexity brought by the pairwise structure. Therefore, we propose the following algorithm to solve for $\hat{\boldsymbol{\beta}}$ and $\hat{\lambda}_k$ ($k = 1, \ldots, m$) iteratively:

Step 1. Start with proper initial values for the parameters, $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$.

Step 2. At the r-th iteration, update each $\lambda_k^{(r)}$ using

$$\lambda_k^{(r)} = \frac{n^{-1}\sum_{i=1}^n I(X_i = w_k)}{n^{-1}\sum_{i=1}^n Y_i(w_k) e^{\mathbf{Z}_i^T \beta^{(r-1)}} + \{n(n-1)\}^{-1} \sum_{i \neq j} \frac{Q_{ij}^{(0)}(w_k; \beta^{(r-1)})}{1 + 1/R_{ij}(\beta^{(r-1)}, \boldsymbol{\lambda}^{(r-1)})}}. \qquad (5)$$

Step 3. Update $\boldsymbol{\beta}^{(r)}$ by one step of Newton-Raphson iteration:

$$\boldsymbol{\beta}^{(r)} = \boldsymbol{\beta}^{(r-1)} - \left(\dot{U}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}^{(r-1)}, \boldsymbol{\lambda}^{(r)})\right)^{-1} \left(U_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{(r-1)}, \boldsymbol{\lambda}^{(r)})\right),$$

where $\dot{U}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}^{(r-1)}, \boldsymbol{\lambda}^{(r)}) = \partial U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\lambda})/\partial \boldsymbol{\beta}^T \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(r-1)}, \boldsymbol{\lambda}=\boldsymbol{\lambda}^{(r)}}$.

Step 4. Repeat Step 2 and 3 until the algorithm converges.

Estimates from the conditional approach can be used as initial values for the parameters in Step 1. Setting $\boldsymbol{\beta}^{(0)} = 0$ and $\boldsymbol{\lambda}^{(0)} = (1/m, \ldots, 1/m)$ is a more convenient alternative. In our simulation studies, it is demonstrated that the algorithm is robust to the choice of initial values. In Step 2, updating λ_k using the self-consistent solution (5) is the crucial step which makes the computation of the PLAC estimator tractable in a reasonable amount of time. The R code implementing the algorithm can be obtained from the authors upon request.

2.3 Asymptotic Properties

In this subsection, we establish the consistency and asymptotic normality of the PLAC estimator $(\hat{\boldsymbol{\beta}}, \hat{\Lambda})$, utilizing techniques from both empirical process (van der Vaart and Wellner, 1996) and U-process theories (De la Peña and Giné, 1999). Denote the normalized score functions corresponding to $\mathcal{L}_n^C$ and $\mathcal{L}_n^P$ as $U^C(\boldsymbol{\beta}, \boldsymbol{\lambda}) = n^{-1}\sum_{i=1}^n U_i^C(\boldsymbol{\beta}, \Lambda)$ and $U^P(\boldsymbol{\beta}, \boldsymbol{\lambda}) = 2\{n(n-1)\}^{-1}\sum_{i<j} U_{ij}^P(\boldsymbol{\beta}, \boldsymbol{\lambda})$, respectively, where

$$U_i^C(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \begin{pmatrix} \Delta_i \mathbf{Z}_i - \mathbf{Z}_i e^{Z_i^T \beta} \sum_{k=1}^m \lambda_k Y_i(w_k) \\ I(X_i = w_1)\{\Delta_i/\lambda_1 - Y_i(w_1)\, e^{\mathbf{Z}_i^T \beta}\} \\ \vdots \\ I(X_i = w_m)\{\Delta_i/\lambda_m - Y_i(w_m)\, e^{\mathbf{Z}_i^T \beta}\} \end{pmatrix} \qquad (6)$$

and

$$
U_{ij}^{P}(\boldsymbol{\beta}, \boldsymbol{\lambda}) \;=\; -1/(1+R_{ij}^{-1}) \begin{pmatrix} \sum_{k=1}^{m} \lambda_k Q_{ij}^{(1)}(w_k) \\ Q_{ij}^{(0)}(w_1) \\ \vdots \\ Q_{ij}^{(0)}(w_m) \end{pmatrix}. \tag{7}
$$

Theorem 1 (Consistency)**.** *Under Conditions (C1)-(C4),*

$$
\hat{\boldsymbol{\beta}} \overset{a.s.}{\to} \boldsymbol{\beta}_0 \quad \textit{and} \quad \left\| \hat{\Lambda} - \Lambda_0 \right\|_{L_\infty[0,\tau]} \overset{a.s.}{\to} 0 \qquad \textit{as } n \to \infty,
$$

where $\|\cdot\|_{L_\infty[0,\tau]}$ *is the supreme norm on* $[0, \tau]$.

Under the regularity conditions specified in the Appendix, Theorem 1 shows that the PLAC estimator is a consistent estimator of the true parameters $(\boldsymbol{\beta}_0, \Lambda_0)$. The consistency proof follows three major steps. First, we show the parameters of interest $(\boldsymbol{\beta}_0, \Lambda_0)$ are identifiable. By the nature of the pairwise construction, $U_{ij}^{P}(\boldsymbol{\beta}, \Lambda)$ is permutation-symmetric in the observed data; thus, the pairwise score function $U^{P}(\boldsymbol{\beta}, \Lambda)$ and its derivatives are U-processes of order two. Second, we construct upper bounds for bracketing numbers of the related function classes by combining the bracketing entropy results of uniformly bounded monotone functions with the preservation theorems for Lipschitz function classes (see van der Vaart and Wellner, 1996, Chapter 2.7). The law of large numbers of these classes then follows from Corollary 3.2.5 of De la Peña and Giné (1999). In addition, we can show $\mathrm{E}\{U^{P}(\boldsymbol{\beta}_0, \Lambda_0)\} = 0$ by the fact that $U_{ij}^{P}(\boldsymbol{\beta}, \Lambda)$ is the exact score function corresponding to the pairwise likelihood of the pair (i, j), conditioning on $(\mathbf{Z}_i, \mathbf{Z}_j)$ and the order statistic of (A_i, A_j). In the last step, the strong consistency of the PLAC estimator can be proven through the likelihood equation argument similar to that given by Murphy et al. (1997), along with the composite Kullback-Leibler divergence (Varin and Vidoni, 2005) and the identifiability of the parameters.

For the weak convergence, we first establish the uniform $\sqrt{n}$-convergence rate and the asymptotic normality of the log-generalized odds ratio using the Hájek projection of U-processes (van der Vaart, 2000). The asymptotic normality of the PLAC estimator can be proved using Theorem 3.3.1 of van der Vaart and Wellner (1996). Noting that $\sqrt{n}\,U(\boldsymbol{\beta}_0, \Lambda_0) = \sqrt{n}\,U^C(\boldsymbol{\beta}_0, \Lambda_0) + \sqrt{n}\,U^P(\boldsymbol{\beta}_0, \Lambda_0)$, the asymptotic normality of $\sqrt{n}\,U(\boldsymbol{\beta}_0, \Lambda_0)$ is obtained by the separate contributions of $\sqrt{n}\,U^C(\boldsymbol{\beta}_0, \Lambda_0)$ and $\sqrt{n}\,U^P(\boldsymbol{\beta}_0, \Lambda_0)$, which are asymptotically independent (van der Vaart and Wellner, 1996, Example 1.4.6). The asymptotic normality of $\sqrt{n}\,U^C(\boldsymbol{\beta}_0, \Lambda_0)$ follows from the martingale theory (Andersen and Gill, 1982; Wang et al., 1993), and our innovative contribution is to identify the limiting distribution of $\sqrt{n}\,U^P(\boldsymbol{\beta}_0, \Lambda_0)$. The normality of the function classes involved in $U^P(\boldsymbol{\beta}_0,\, \Lambda_0)$ and its derivative is shown through the results on the VC subgraph classes, the normality of the log-generalized odds ratio, and the preservation theorems for Lipschitz functions (van der Vaart and Wellner, 1996, Chapter 2.10). Finally, the Fréchet-differentiability of $\mathrm{E}\{U(\boldsymbol{\beta}_0,\, \Lambda_0)\}$ and the invertibility of its derivative can be shown by (C5) and the Fredholm theory, following arguments similar to those in Zeng and Lin (2006). Further detailed proofs of Theorems 1-2 are provided in the Appendix.

One of the appealing features of our approach is that the covariance function of the limiting process of the PLAC estimator can be consistently estimated by a closed-form sandwich estimator. To define the asymptotic covariance function, consider a linear functional

$$\sqrt{n}\left\{b_1^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \int_0^\tau h(t)d\left(\hat{\Lambda}(t) - \Lambda_0(t)\right)\right\}, \tag{8}$$

where b_1 is a vector in $\mathbb{R}^p$, $h(t)$ is an arbitrary function with bounded total variation on $[0, \tau]$. Let b_2 be the $m \times 1$ vector $(h(w_1), \ldots, h(w_m))^T$, and $b^T = (b_1^T, b_2^T)$. We further define

$$\begin{aligned}
\hat{V}^C &= \frac{1}{n}\sum_{i=1}^n U_i^C(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}})^{\otimes 2}, \\
\hat{J}^C &= -\frac{1}{n}\sum_{i=1}^n \partial U_i^C(\boldsymbol{\beta}, \boldsymbol{\lambda})/\partial(\boldsymbol{\beta}^T, \boldsymbol{\lambda}^T)\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\, \boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}}, \\
\hat{V}^P &= \frac{4}{n-1}\sum_{i=1}^n \left\{\frac{1}{n-1}\sum_{i\neq j} U_{ij}^P(\hat{\boldsymbol{\beta}},\, \hat{\boldsymbol{\lambda}})\right\}^{\otimes 2}, \\
\hat{J}^P &= -\frac{1}{n(n-1)}\sum\sum_{i\neq j} \partial U_{ij}^P(\boldsymbol{\beta}, \boldsymbol{\lambda})/\partial(\boldsymbol{\beta}^T, \boldsymbol{\lambda}^T)\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\, \boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}},
\end{aligned}$$

where the exact expressions of $\partial U_i^C(\boldsymbol{\beta}, \boldsymbol{\lambda})/\partial(\boldsymbol{\beta}^T, \boldsymbol{\lambda}^T)$ and $\partial U_{ij}^P(\boldsymbol{\beta}, \boldsymbol{\lambda})/\partial(\boldsymbol{\beta}^T, \boldsymbol{\lambda}^T)$ are given in the web supplementary materials. As in Zeng and Lin (2006), since the PLAC estimator for Λ converges at a parametric rate, we can treat $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ in (3) as if they are finite-dimensional parameters. Then by the asymptotic properties of U-statistics (Sen, 1960) and the composite likelihood theory, we can estimate the asymptotic covariance function by the inverse of the observed Godambe information matrix (Varin et al., 2011). The weak convergence results are summarized in the following theorem.

Theorem 2 (Asymptotic normality)**.** *Under Conditions (C1)-(C5),* $\sqrt{n}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0,\, \hat{\Lambda}(t) - \Lambda_0(t))$ *converges weakly to a mean-zero Gaussian process in* $\mathbb{R}^p \times \mathrm{BV}[0,\tau]$*, where* $\mathrm{BV}[0,\tau]$ *denotes the space of all functions with bounded total variations on* $[0,\tau]$*. In addition, the linear functional* (8) *converges in distribuion to a mean-zero Gaussian random variable with the variance that can be consistently estimated by* $b^T\hat{\boldsymbol{\Sigma}}b$*, where*

$$\hat{\boldsymbol{\Sigma}} = (\hat{J}^C + \hat{J}^P)^{-1}(\hat{V}^C + \hat{V}^P)(\hat{J}^C + \hat{J}^P)^{-1}. \tag{9}$$

Naturally, the estimated asymptotic covariance matrix (9) has the following partition:

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}\boldsymbol{\lambda}} \\ \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\lambda}\boldsymbol{\beta}} & \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\lambda}\boldsymbol{\lambda}} \end{pmatrix},$$

where the sub-matrices correspond to the estimated asymptotic variance-covariance matrices of the corresponding parameter estimates. Recall that $\hat{\Lambda}(t) = \sum_{k=1}^m \hat{\lambda}_k I(w_k \leqslant t)$. From Theorem 2, the asymptotic variance of $\hat{\Lambda}(t)$, denoted by $\boldsymbol{\Sigma}_{\hat{\Lambda}(t)}$, can be estimated by setting $b_1 = 0$ and $b_2^T = (I(w_1 \leqslant t), \ldots, I(w_m \leqslant t))$, i.e.,

$$\hat{\boldsymbol{\Sigma}}_{\hat{\Lambda}(t)} = \sum_{k=1}^m \sum_{l=1}^m I(w_k \leqslant t, w_l \leqslant t)\hat{\sigma}_{kl}^{(\boldsymbol{\lambda\lambda})},$$

where $\hat{\sigma}_{kl}^{(\boldsymbol{\lambda\lambda})}$ is the covariance estimate corresponding to λ_k and λ_l in the sub-matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\lambda\lambda}}$.

The greatest advantage of having a closed-form variance estimator (9) is that the asymptotic variances of other quantities of interest can be estimated directly by the delta method. For example, one may be interested in estimating

$$S_{\mathbf{z}_0}(t) = \exp\left(-\Lambda_{\mathbf{z}_0}(t)\right) = \exp\left(-\Lambda(t)\exp(\mathbf{z}_0^T\boldsymbol{\beta})\right),$$

i.e., the survival probability of an individual with covariates $\mathbf{z}_0$ along with its point-wise confidence interval (CI). Here, we use the log-log transformed confidence interval (Borgan and Liestøl, 1990) to avoid CIs with ranges outside of $[0,1]$. Plugging in the PLAC estimator, at a fixed time t, $\sqrt{n}\left(\log\hat{\Lambda}_{\mathbf{z}_0}(t) - \log\Lambda_{\mathbf{z}_0}(t)\right)$ is asymptotically normal. Using the delta method, we can estimate its variance by

$$\hat{\mathbf{\Sigma}}_{\log\hat{\Lambda}_{\mathbf{z}_0}(t)} = \begin{pmatrix}\mathbf{z}_0^T & \hat{\Lambda}^{-1}(t)\end{pmatrix}\begin{pmatrix}\hat{\mathbf{\Sigma}}_{\beta\beta} & \hat{\mathbf{\Sigma}}_{\hat{\beta},\hat{\Lambda}(t)} \\ \hat{\mathbf{\Sigma}}^T_{\hat{\beta},\hat{\Lambda}(t)} & \hat{\mathbf{\Sigma}}_{\hat{\Lambda}(t)}\end{pmatrix}\begin{pmatrix}\mathbf{z}_0 \\ \hat{\Lambda}^{-1}(t)\end{pmatrix},$$

where $\mathbf{\Sigma}_{\hat{\beta},\hat{\Lambda}(t)}$ denotes the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ and $\hat{\Lambda}(t)$, which can be obtained similarly as above by summing up the corresponding elements of (9). A $100\times(1-\alpha)\%$ CI for $\log\Lambda_{\mathbf{z}_0}(t)$ is given by $(\log\hat{\Lambda}_{\mathbf{z}_0}(t) \pm z_{1-\alpha/2}\hat{\mathbf{\Sigma}}^{1/2}_{\log\hat{\Lambda}_{\mathbf{z}_0}(t)}/\sqrt{n}))$, where $z_{1-\alpha/2}$ denotes the upper $(1-\alpha/2)$ quantile of the standard normal distribution. Let $\zeta = \exp(z_{1-\alpha/2}\hat{\mathbf{\Sigma}}^{1/2}_{\log\hat{\Lambda}_{\mathbf{z}_0}(t)}/\sqrt{n})$. By taking inverse of the log-log transformation of the upper and lower bounds of the CI for $\log\Lambda_{\mathbf{z}_0}(t)$, a $100\times(1-\alpha)\%$ CI for $S_{\mathbf{z}_0}(t)$ is given by

$$\left(\hat{S}^{1/\zeta}_{\mathbf{z}_0}(t), \hat{S}^{\zeta}_{\mathbf{z}_0}(t)\right), \quad t\in[0,\tau].$$

3 Simulation Study

We conducted extensive simulation studies to evaluate the finite-sample performance of the proposed PLAC estimator, and compared it with estimators using the conditional approach (Conditional) by Wang et al. (1993), and the composite partial likelihood method (CPL) by Huang and Qin (2012). The underlying survival time T^* was generated from a Cox model with two independent covariates:

$$\lambda(t|Z_1, Z_2; \beta_1, \beta_2) = \lambda(t)\exp(\beta_1 Z_1 + \beta_2 Z_2), \tag{10}$$

where $Z_1 \sim$ Binomial(0.5) and $Z_2 \sim$ Uniform$[-1,1]$; and the true values of β_1 and β_2 were set to be 1. The baseline hazard function was $\lambda(t) = 2t$, which corresponded to a Weibull distribution. For the underlying truncation time, we considered two cases:

Case 1. Length-biased sampling data with $A^* \sim$ Uniform$[0, \tau]$;

Case 2. Non-length-biased sampling data with $A^* \sim$ Exponential(1).

In Case 1, we first generated observed survival times t_i, $i = 1, \ldots, n$, and then drew corresponding truncation times a_i from Uniform$[0, t_i]$ (see, e.g., Mandel and Betensky, 2007). In Case 2, the underlying survival times t_i^* were generated from (10), and the underlying truncation times were from Exponential(1); yet only the pairs (a_i^*, t_i^*) satisfying $a_i^* < t_i^*$ were kept until the desired sample size was reached. The censoring times c_i, $i = 1, \ldots, n$, were generated from Uniform$[0, C_{\max}]$ independently, where $C_{\max}$ was chosen to designate various censoring rates of approximately 20%, 50% and 80%. The censoring indicators for subject i was obtained by $\delta_i = I(t_i \leqslant a_i + c_i)$. Sample sizes of 200, 400 and 800 were considered. We generated 1000 datasets under each scenario.

For each simulated dataset, we estimated β_1, β_2, and $\Lambda(t)$ at two fixed time points $t = (\tau_{30}, \tau_{60})$, where τ_{30} and τ_{60} were the 30% and 60% percentiles of the observed survival times under each scenario. Summary statistics, including the average of the esimates minus the true value (Bias), the empirical standard error of the estimates (SE), the average of the standard error estimates (SEE), the 95% coverage probability (CP), the mean square error (MSE) and the asymptotic relative efficiency (ARE) compared to the conditional estimator, are provided in Table 1 (for Case 1) and Table 2 (for Case 2).

The empirical biases of the PLAC estimates, like the conditional approach estimates, are close to zero under all scenarios, especially when the sample size raises to 400 or 800. When the censoring rate is high (PC=80%), the biases of the PLAC estimates are a bit larger than those of the conditional approach estimates, which is possibly due to the common bias-variance trade-off. When data are length-biased, the CPL estimator also enjoys close-to-zero biases (see results for Case 1). In contrast, the CPL estimates in Case 2 (non-length-biased data) are severely biased, and the biases remain at similar magnitudes even when the sample size increases.

The PLAC estimator yields considerable efficiency gains compared to the corresponding conditional approach estimator. Under different sample sizes and censoring rates, the efficiency gains in $\hat{\beta}_1$ and $\hat{\beta}_2$ range from 35% to 165% in Case 1 and 22% to 96% in Case 2. The efficiency gains in $\hat{\Lambda}_{\tau_{30}}$ and $\hat{\Lambda}_{\tau_{60}}$ are not as significant, but improvement over the

n	PC		Conditional			CPL				PLAC					
			Bias	SE	MSE	Bias	SE	MSE	ARE	Bias	SE	SEE	CP	MSE	ARE
200	20	$\hat{\beta}_1$	.010	.187	.035	.003	.158	.025	1.40	.007	.153	.151	94.7	.024	1.48
		$\hat{\beta}_2$	.009	.169	.029	.005	.141	.020	1.44	.008	.140	.137	94.7	.020	1.46
		$\hat{\Lambda}_{\tau_{30}}$	-.001	.065	.004	-.003	.058	.003	1.26	-.002	.061	.060	93.2	.004	1.15
		$\hat{\Lambda}_{\tau_{60}}$	.004	.133	.018	.000	.116	.013	1.31	.002	.125	.122	93.7	.016	1.13
	50	$\hat{\beta}_1$	.016	.231	.054	.006	.195	.038	1.41	.015	.179	.178	95.2	.032	1.66
		$\hat{\beta}_2$	.008	.209	.044	.005	.176	.031	1.42	.013	.165	.160	94.8	.027	1.61
		$\hat{\Lambda}_{\tau_{30}}$	-.001	.060	.004	-.002	.055	.003	1.22	-.003	.056	.054	91.1	.003	1.18
		$\hat{\Lambda}_{\tau_{60}}$	-.003	.118	.014	-.004	.106	.011	1.25	-.006	.108	.110	93.7	.012	1.19
	80	$\hat{\beta}_1$	.038	.391	.154	.009	.311	.097	1.58	.044	.255	.239	95.0	.067	2.34
		$\hat{\beta}_2$	.034	.362	.132	.012	.288	.083	1.58	.049	.232	.218	93.8	.056	2.43
		$\hat{\Lambda}_{\tau_{30}}$	.000	.055	.003	-.001	.048	.002	1.28	-.003	.046	.042	88.3	.002	1.41
		$\hat{\Lambda}_{\tau_{60}}$	-.004	.128	.016	-.007	.110	.012	1.36	-.011	.109	.105	90.6	.012	1.38
400	20	$\hat{\beta}_1$	-.003	.129	.017	-.004	.105	.011	1.50	-.003	.105	.106	94.6	.011	1.52
		$\hat{\beta}_2$	-.004	.110	.012	-.004	.095	.009	1.33	-.004	.094	.096	94.9	.009	1.37
		$\hat{\Lambda}_{\tau_{30}}$	.002	.047	.002	.000	.041	.002	1.31	.001	.043	.043	93.8	.002	1.15
		$\hat{\Lambda}_{\tau_{60}}$	.008	.094	.009	.004	.081	.007	1.33	.007	.087	.087	94.9	.008	1.15
	50	$\hat{\beta}_1$	.000	.160	.026	-.003	.126	.016	1.60	.002	.122	.124	94.9	.015	1.72
		$\hat{\beta}_2$	-.001	.144	.021	.000	.117	.014	1.50	.002	.113	.113	94.3	.013	1.61
		$\hat{\Lambda}_{\tau_{30}}$	.002	.042	.002	.000	.037	.001	1.33	.000	.039	.039	95.2	.001	1.19
		$\hat{\Lambda}_{\tau_{60}}$	.004	.086	.007	.001	.075	.006	1.32	.001	.079	.079	94.4	.006	1.21
	80	$\hat{\beta}_1$	.016	.273	.075	.000	.208	.043	1.72	.023	.168	.165	94.1	.029	2.65
		$\hat{\beta}_2$	.008	.235	.055	.001	.184	.034	1.64	.022	.156	.152	95.1	.025	2.26
		$\hat{\Lambda}_{\tau_{30}}$	.001	.037	.001	.000	.033	.001	1.27	-.002	.032	.031	91.4	.001	1.37
		$\hat{\Lambda}_{\tau_{60}}$	-.001	.092	.008	-.002	.079	.006	1.36	-.007	.078	.076	91.2	.006	1.41
800	20	$\hat{\beta}_1$	-.001	.090	.008	-.002	.075	.006	1.41	-.001	.075	.075	94.6	.006	1.42
		$\hat{\beta}_2$	-.001	.080	.006	-.001	.070	.005	1.31	-.001	.069	.068	94.6	.005	1.35
		$\hat{\Lambda}_{\tau_{30}}$	.000	.033	.001	.000	.030	.001	1.24	.000	.032	.030	94.9	.001	1.09
		$\hat{\Lambda}_{\tau_{60}}$	.002	.064	.004	.001	.057	.003	1.28	.001	.061	.061	94.9	.004	1.11
	50	$\hat{\beta}_1$	-.002	.114	.013	-.004	.094	.009	1.48	-.001	.091	.088	94.3	.008	1.59
		$\hat{\beta}_2$	-.003	.099	.010	-.003	.083	.007	1.42	-.001	.080	.080	94.5	.006	1.55
		$\hat{\Lambda}_{\tau_{30}}$	.001	.031	.001	.000	.028	.001	1.25	.000	.029	.028	93.7	.001	1.15
		$\hat{\Lambda}_{\tau_{60}}$	.004	.061	.004	.002	.054	.003	1.30	.002	.057	.056	95.0	.003	1.16
	80	$\hat{\beta}_1$	.004	.183	.034	-.001	.142	.020	1.66	.009	.121	.116	94.1	.015	2.31
		$\hat{\beta}_2$	-.004	.163	.027	-.006	.127	.016	1.63	.006	.110	.106	94.2	.012	2.20
		$\hat{\Lambda}_{\tau_{30}}$	.001	.025	.001	.000	.023	.001	1.25	.000	.022	.022	93.1	.001	1.28
		$\hat{\Lambda}_{\tau_{60}}$	.001	.064	.004	-.001	.055	.003	1.33	-.002	.056	.054	93.8	.003	1.30

Table 1: Summaries of 1000 simulations in Case 1 (length-biased data). PC: proportion of censoring; Bias, SE, SEE, CP and MSE: empirical bias, standard error, SE estimate, 95% coverage probability and mean square error; ARE: asymptotic relative efficiency with respect to the conditional approach estimator. The true values for β_1 and β_2 are 1, and $\hat{\Lambda}(t)$ are evaluated at the 30% and 60% percentiles (τ_{30} and τ_{60}) of the observed survival times.

n	PC		Conditional			CPL				PLAC					
			Bias	SE	MSE	Bias	SE	MSE	ARE	Bias	SE	SEE	CP	MSE	ARE
200	20	$\hat{\beta}_1$	.003	.176	.031	-.090	.149	.030	1.40	.004	.157	.157	94.4	.025	1.26
		$\hat{\beta}_2$	.006	.165	.027	-.088	.138	.027	1.43	.008	.147	.142	94.6	.022	1.26
		$\hat{\Lambda}_{\tau_{30}}$	-.003	.062	.004	.072	.070	.010	.78	-.004	.059	.057	91.3	.004	1.08
		$\hat{\Lambda}_{\tau_{60}}$	-.003	.114	.013	.165	.123	.042	.86	-.005	.110	.109	94.6	.012	1.08
	50	$\hat{\beta}_1$	.000	.226	.051	-.095	.185	.043	1.49	.006	.190	.186	94.5	.036	1.42
		$\hat{\beta}_2$	.003	.210	.044	-.092	.175	.039	1.43	.009	.178	.167	93.5	.032	1.39
		$\hat{\Lambda}_{\tau_{30}}$	-.001	.056	.003	.053	.063	.007	.77	-.003	.053	.050	90.3	.003	1.12
		$\hat{\Lambda}_{\tau_{60}}$	.000	.110	.012	.125	.119	.030	.86	-.003	.103	.098	92.7	.011	1.14
	80	$\hat{\beta}_1$	.020	.388	.151	-.080	.303	.098	1.64	.051	.277	.258	93.8	.079	1.96
		$\hat{\beta}_2$	.016	.341	.117	-.085	.271	.081	1.58	.038	.254	.232	94.4	.066	1.81
		$\hat{\Lambda}_{\tau_{30}}$	.002	.045	.002	.029	.052	.003	.76	-.002	.040	.038	89.8	.002	1.27
		$\hat{\Lambda}_{\tau_{60}}$	.002	.097	.009	.068	.105	.016	.86	-.006	.086	.081	90.6	.007	1.28
400	20	$\hat{\beta}_1$	.002	.130	.017	-.090	.108	.020	1.43	.003	.115	.111	93.3	.013	1.27
		$\hat{\beta}_2$	-.002	.115	.013	-.096	.097	.019	1.42	-.002	.104	.100	94.7	.011	1.24
		$\hat{\Lambda}_{\tau_{30}}$	.000	.043	.002	.076	.050	.008	.76	-.001	.042	.040	93.0	.002	1.04
		$\hat{\Lambda}_{\tau_{60}}$	-.002	.083	.007	.168	.089	.036	.87	-.003	.080	.077	93.3	.006	1.06
	50	$\hat{\beta}_1$	.002	.162	.026	-.090	.129	.025	1.57	.006	.136	.131	94.0	.019	1.42
		$\hat{\beta}_2$	.000	.141	.020	-.093	.119	.023	1.40	.002	.122	.118	94.3	.015	1.33
		$\hat{\Lambda}_{\tau_{30}}$	.000	.040	.002	.055	.046	.005	.76	-.001	.038	.035	91.4	.001	1.09
		$\hat{\Lambda}_{\tau_{60}}$	-.001	.076	.006	.125	.083	.023	.84	-.003	.073	.069	93.5	.005	1.09
	80	$\hat{\beta}_1$	.003	.254	.065	-.091	.200	.048	1.62	.020	.185	.178	94.5	.035	1.88
		$\hat{\beta}_2$	.012	.230	.053	-.087	.186	.042	1.52	.019	.173	.162	93.4	.030	1.77
		$\hat{\Lambda}_{\tau_{30}}$	.000	.032	.001	.028	.037	.002	.74	-.002	.029	.027	90.3	.001	1.21
		$\hat{\Lambda}_{\tau_{60}}$	-.001	.067	.004	.067	.073	.010	.83	-.005	.061	.058	91.4	.004	1.21
800	20	$\hat{\beta}_1$	.005	.088	.008	-.088	.073	.013	1.43	.005	.079	.079	95.0	.006	1.23
		$\hat{\beta}_2$	.000	.078	.006	-.094	.066	.013	1.40	.001	.070	.071	95.0	.005	1.22
		$\hat{\Lambda}_{\tau_{30}}$	-.001	.029	.001	.077	.033	.007	.74	-.001	.028	.029	94.9	.001	1.04
		$\hat{\Lambda}_{\tau_{60}}$	-.002	.054	.003	.170	.059	.032	.84	-.002	.053	.055	95.1	.003	1.04
	50	$\hat{\beta}_1$	.006	.111	.012	-.089	.091	.016	1.48	.007	.094	.092	95.1	.009	1.41
		$\hat{\beta}_2$	-.004	.097	.009	-.094	.082	.016	1.40	-.001	.084	.083	94.8	.007	1.33
		$\hat{\Lambda}_{\tau_{30}}$	-.001	.027	.001	.056	.031	.004	.72	-.001	.026	.025	93.6	.001	1.07
		$\hat{\Lambda}_{\tau_{60}}$	-.001	.051	.003	.128	.057	.020	.80	-.002	.049	.049	94.4	.002	1.08
	80	$\hat{\beta}_1$	.010	.181	.033	-.085	.140	.027	1.67	.015	.133	.126	94.4	.018	1.84
		$\hat{\beta}_2$	-.002	.155	.024	-.094	.125	.025	1.54	.006	.118	.114	95.1	.014	1.72
		$\hat{\Lambda}_{\tau_{30}}$	.000	.022	.000	.028	.026	.001	.72	-.001	.021	.020	92.9	.000	1.16
		$\hat{\Lambda}_{\tau_{60}}$	-.001	.047	.002	.069	.053	.008	.79	-.002	.044	.041	93.0	.002	1.17

Table 2: Summaries of 1000 simulations in Case 2 (non-length-biased data). PC: proportion of censoring; Bias, SE, SEE, CP and MSE: empirical bias, standard error, SE estimate, 95% coverage probability and mean square error; ARE: asymptotic relative efficiency with respect to the conditional approach estimator. The true values for β_1 and β_2 are 1, and $\hat{\Lambda}(t)$ are evaluated at the 30% and 60% percentiles (τ_{30} and τ_{60}) of the observed survival times.

conditional approach has been shown, i.e., all AREs are greater than one. In Case 1, we observe that the PLAC estimator of the regression coefficients is even more precise than the CPL estimator that aims at improving efficiency for the length-biased sampling cases. This can be partly explained by the fact that the CPL method is also not fully efficient. Note that the relative efficiency gains of the PLAC estimator increase as the censoring rate increases, because the augmenting pairwise-likelihood is not subject to censoring. Taking the biases and the variances altogether, the MSE of the PLAC estimator are either the smallest, or comparable to those of the other two estimators.

Comparing the SEs and SEEs of the PLAC estimator, we demonstrate that the variance of the PLAC estimator is consistently estimated by the proposed variance estimator (9). We notice that the SEs for the PLAC estimates under $n = 800$ are approximately a half of the corresponding SEs under $n = 200$, which confirms the $\sqrt{n}$-convergence rate of the PLAC estimator as proven in Section 2.3. In the scenario with $n = 200$ and 80% censoring rate, the 95% CPs for the PLAC estimator are close to the nominal level, except for $\hat{\Lambda}_{\tau_{30}}$ and $\hat{\Lambda}_{\tau_{60}}$. This is not due to the method but because of the small number of observed events that attenuates the normal approximation not only in our approach, but also in others. For example, under 80% censoring, the CPs of $\hat{\Lambda}_{\tau_{30}}$ and $\hat{\Lambda}_{\tau_{60}}$ using the conditional approach are 87.8% and 91.1%, both of which are also under the nominal level. Nevertheless, when the sample size increases to 400 and 800, the 95% CPs of the PLAC estimator get closer to the nominal level.

In summary, the proposed PLAC estimator performs well under finite sample sizes. It is unbiased, and enjoys substantial gains in efficiency in both the regression coefficients and the cumulative baseline hazard function. The performance of PLAC is robust to the violation of the stationarity assumption as well as high censoring rates. The proposed sandwich estimator results in good variance estimates for all parameters, and yields resonable confidence intervals.

We further performed addtional simulations with the baseline hazard function $\lambda(t) = 1$. The results (not shown here) were as good as in Table 1-2 or even better with slightly increased efficiency gains.

4 Data Application

We now apply the proposed method to the RRI-CKD study introduced in Section 1. In this study, the survival time was measured from the referral to the composite renal outcome defined as either death, long-term dialysis or kidney transplantation, whichever came first. The truncation time was measured from the referral to the study enrollment. Patients without referral information were excluded. A total of 775 patients were included in our analysis, among which 364 experienced the composite renal outcome during the study follow-up, and the censoring rate was 53%. The baseline patient characteristics included demographics (age, gender, race), estimated GFR (eGFR), co-morbidities (hypertension and diabetes) and centers (University of Michigan, University of North Carolina, Yale University or Albany Medical College).

We first assessed whether the referrals occurred at a constant rate (i.e., the stationarity assumption) using the fact that the truncation time A should follow the same distribution as the residual survival time V when the stationarity assumption is satisfied. We conducted a log-rank test for the pair (A, V) (Jung, 1999; Mandel and Betensky, 2007), and the null hypothesis of the same distribution was rejected with $p < 0.001$. The significant result was confirmed by the graphical checking method by Asgharian et al. (2006).

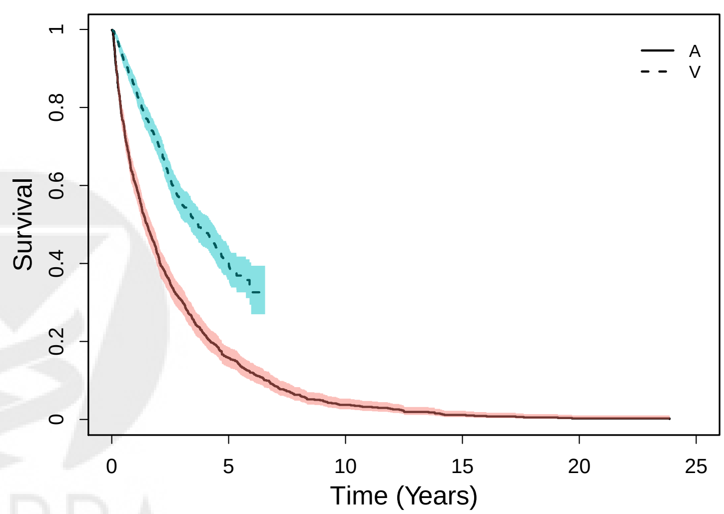

Figure 2: Kaplan-Meier curves for the truncation time A (solid) and the residual survival time V (dashed) of the RRI-CKD data. The 95% point-wise CIs are shown as shaded areas.

As shown in Figure 2, the Kaplan-Meier curve of V is above that of A throughout, and their point-wise confidence intervals do not overlap. Therefore, we concluded that the stationarity assumption did not hold in the RRI-CKD data, and hence length-biased-sampling-specific methods like the CPL estimator might yield invalid inference. The violation of the stationarity assumption can be partly explained by the absence of general guidelines for when to refer to a nephrologist in practice; patients can be referred at either early or late stages of CKD.

We compared analysis results from the proposed method (PLAC) with those from the conditional approach (Conditional). Table 3 lists the estimated regression coefficients of the baseline covariates and their standard errors. A forest plot of the hazards ratios of the covarates is shown in Figure 3 to visualize the significance of the estimates.

Coefficient	Conditional	PLAC
Albany	-.122 (.162)	-.014 (.124)
Yale	-.113 (.147)	-.261 (.117)
UNC	-.217 (.145)	-.023 (.118)
Age	-.096 (.059)	.078 (.048)
Male	.269 (.111)	.234 (.092)
Non-White	.316 (.128)	.378 (.102)
Diabetes	.405 (.115)	.476 (.094)
Hypertension	.208 (.196)	-.010 (.148)
eGFR	-.417 (.090)	-.465 (.077)

Table 3: Estimates (SEs) of the regression coefficients using the conditional approach (Conditional) and the proposed estimator (PLAC) for the RRI-CKD data.

Compared with the conditional approach, the PLAC estimator estimates all coefficients with improved precision (smaller standard errors and narrower 95% confidence intervals); the variance ratio of the conditional approach estimate to the corresponding PLAC estimate is 1.36 or greater. This implies that the conditional approach requires at least 36% more CKD patients to achieve the same estimating precision as the PLAC estimator. The coefficient estimates of the centers, baseline age, and presence of hypertension using the two methods show obvious deviations, among which the difference in age cofficients is significant; although both estimates themselves are not significant.

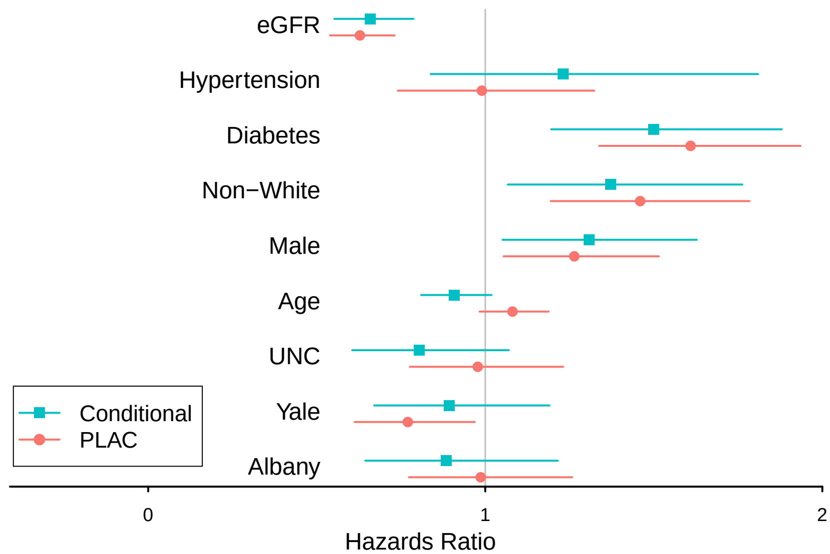

Figure 3: Estimated hazards ratios of the covariates for the RRI-CKD data. The squares and dots represent the estimates using the conditional approach and the proposed method (PLAC). The horizontal lines around the points represent the corresponding 95% CIs.

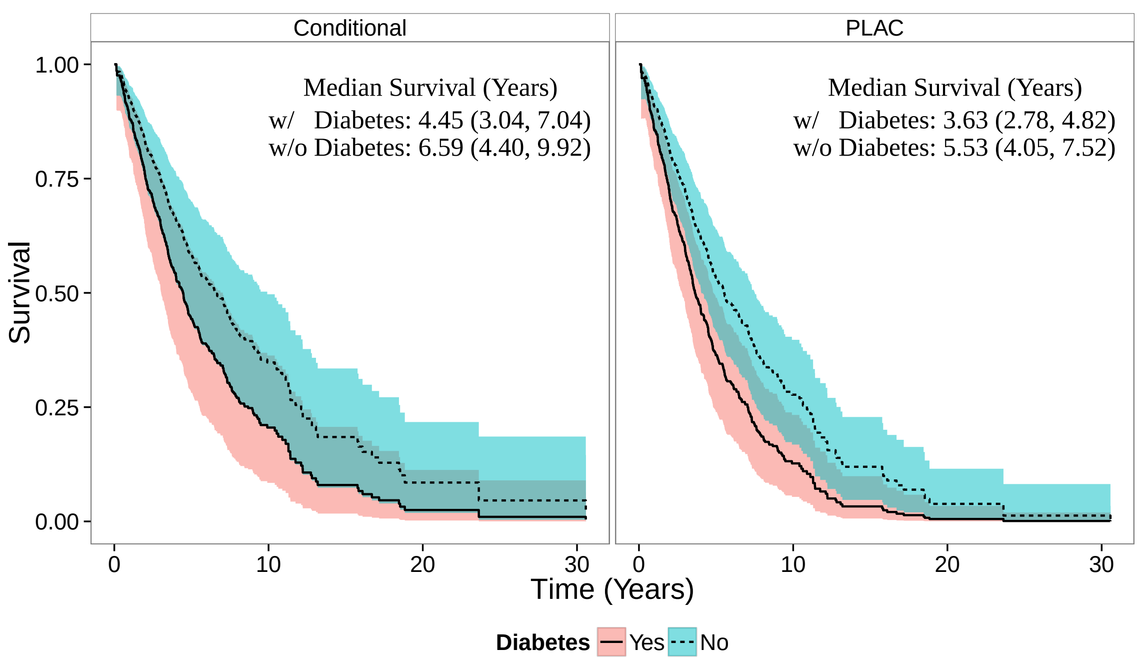

Figure 4: Estimated survival curves of patients with diabetes (solid) or without diabetes (dashed) at baseline using the conditional approach (Conditional) and the proposed method (PLAC). Log-log transformed 95% point-wise CIs are shown as shaded areas. The estimated median survival times for both groups are displayed in each panel with the corresponding 95% CIs. The other covariates are set to their means or the reference levels.

As an example of the interesting application of the PLAC estimator described at the end of Section 2.3, we estimated the survival curves of patients with and without diabetes at baseline, and constructed the corresponding 95% point-wise CIs. The survival estimates and the confidence intervals using the conditional approach and the PLAC estimator are plotted side-by-side in Figure 4. Compared with the conditional approach, the PLAC estimator yields narrower point-wise CIs. We also present the estimated median survival times (and the corresponding 95% CIs), which are the time coordinates of the estimated survival curves (and the 95% point-wise CIs) crossing the horizontal line at 0.5.

5 Discussion

In this manuscript we have proposed a semiparametric estimation method for the Cox model with the issue of general left-truncation. By constructing a pairwise likelihood from the marginal likelihood of the truncation times, we have eliminated the unknown truncation distribution from the full likelihood. Based on our simulation studies, the PLAC estimator has been shown to be robust to heavy censoring rates and the violation of the stationarity assumption. On the contrary, all length-biased sampling methods of efficiency improvement, such as the CPL estimator, rely on the stationarity assumption to lead to consistent estimates. The robustness of the PLAC estimator here means consistency and efficiency gain over the conditional approach estimator across all scenarios we considered. The gain in efficiency is the greatest advantage of the proposed method. Comparing to the conditional approach (Kalbfleisch and Lawless, 1991; Wang et al., 1993), we observed an efficiency gain of 36% or more in the estimates of the regression coefficients in the analysis of the RRI-CKD data. Our simulations show that the PLAC estimator is even more efficient than the CPL estimator under the length-biased sampling scenario. Note that the CPL method (Huang and Qin, 2012) is based on the correctly specified uniform distribution of the truncation times, exploiting the exchangeability of A and V for the uncensored subjects. Their composite partial likelihood takes the form

$$\mathcal{L}_n^{CPL} = \prod_{i=1}^{n} \frac{f(X_i|\mathbf{Z}_i)^{\Delta_i} S(X_i|\mathbf{Z}_i)^{1-\Delta_i}}{S(A_i|\mathbf{Z}_i)} \times \prod_{i=1}^{n} \left\{ \frac{f(X_i|\mathbf{Z}_i)}{S(\tilde{V}_i|\mathbf{Z}_i)} \right\}^{\Delta_i}, \tag{11}$$

where $\tilde{V}_i = \min(V_i, C_i)$. The additional term in (11) compared with the conditional likelihood $\mathcal{L}_n^C$ is $\prod_{i=1}^n \{f(X_i|\mathbf{Z}_i)/S(\tilde{V}_i|\mathbf{Z}_i)\}^{\Delta_i}$. It implies that the extra information used in the CPL estimator is still as a form of the conditional likelihood accounting for the uncensored subjects only, whereas the PLAC estimator uses the information coming from the (pairwise) marginal likelihood accounting for both censored and uncensored subjects. This different extents of extra information usage could be a major reason behind the higher efficiency gain of the PLAC estimator over the CPL estimator. The efficiency loss compared to the full-likelihood approach under the stationarity assumption warrants further research.

We utilized a NPMLE-type estimator to estimate the cumulative baseline hazard function along with the regression coefficients. Under regularity conditions, the consistency and asymptotic normality of $(\hat{\boldsymbol{\beta}}, \hat{\Lambda})$ have been rigorously proved which results in a closed-form consistent sandwich variance estimator. As an alternative to the NPMLEs, we have studied the maximum pseudo-likelihood (MPL) method in Huang et al. (2012), where Λ in the pairwise likelihood $\mathcal{L}_n^P$ is replaced with the Breslow estimator obtained from $\mathcal{L}_n^C$, and the resulting pseudo-likelihood is used to estimate $\boldsymbol{\beta}$. Our simulation studies (not shown here) revealed loss of efficiency in the MPL estimator compared with the PLAC estimator. This is expected because the plug-in-type pseudo-likelihood implies a further approximation to $\mathcal{L}_n^P$, and thus the information of Λ may not be fully used.

While the proposed PLAC estimator focuses on handling time-independent covariates, the extension to time-dependent covariates is promising based on our preliminary work. We expect to derive asymptotic properties and devote more effort to reducing computation time, which is magnified by the need of expanding the dataset with time-dependent covariates.

A Appendix

In this appendix, we prove the asymptotic results of the proposed PLAC estimator under the following regularity conditions. Note that (C4) is necessary for $\mathcal{L}_n^P$ to be non-degenerate so that we can attain efficiency gains beyond the conditional approach. Detailed explanation of the conditions can be found in the web supplementary materials.

(C1) The true regression coefficients vector $\boldsymbol{\beta}_0$ lies in the interior of a compact set $B \subset \mathbb{R}^p$. The true cumulative baseline hazard function $\Lambda_0(t)$ is continuously differentiable and strictly increasing on $[0, \tau]$, and satisfies $\Lambda_0(0) = 0$.

(C2) The covariates vector $\mathbf{Z}$ is bounded almost surely. If there exist a deterministic function $b_0(t)$ and a vector $b \in \mathbb{R}^p$, such that $b_0(t) + b^T\mathbf{Z} = 0$ with probability one, then $b_0(t) = 0$ and $b = 0$.

(C3) With probability one, there exists a constant $\delta_1 > 0$ such that $\Pr(A^* < T^* \leqslant A^* + C \mid \mathbf{Z},\, A^*,\, C) > \delta_1$, $\Pr(A + C \geqslant \tau \mid \mathbf{Z}) > \delta_1$, and that $\Pr(T \geqslant \tau \mid \mathbf{Z}) > \delta_1$.

(C4) With probability one, there exists a constant $\delta_2 > 0$ such that $\Pr(A^* \geqslant T^* | \mathbf{Z}) > \delta_2$.

(C5) Let $b \in \mathbb{R}^p$, and h be a function with bounded total variation on $[0, \tau]$, then the information operator corresponding to the conditional likelihood evaluated at the true parameters $(\boldsymbol{\beta}_0,\, \Lambda_0)$, $J_0^C(b,\, h) = \left(\lim_{n\to\infty} \partial U^C(\boldsymbol{\beta}, \Lambda)/\partial(\boldsymbol{\beta}, \Lambda)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0,\, \Lambda=\Lambda_0} \right)(b,\, h)$ is invertible.

We use Ω to denote the set of all possible observations. For convenience, we adopt the de Finetti's linear functional notations (Pollard, 2002), where $\mathbb{P}_n$ denotes the empirical measure of the observations $\mathcal{O}_i$, $i = 1, \ldots, n$, P_0 denotes the true probability measure on Ω, and $\mathbb{U}_{n,2}$ is the empirical measure of pairs $(\mathcal{O}_i, \mathcal{O}_j)$ such that $1 \leqslant i < j \leqslant n$.

A.1 Identifiability of $(\boldsymbol{\beta}_0, \Lambda_0)$

Lemma 1. *Under Conditions (C1)-(C3), both $\boldsymbol{\beta}_0$ and Λ_0 are identifiable. Specifically, if there exist parameters $(\boldsymbol{\beta},\, \Lambda)$ such that Λ is absolutely continuous with respect to Λ_0, $\ell_n^C(\boldsymbol{\beta},\, \Lambda) = \ell_n^C(\boldsymbol{\beta}_0,\, \Lambda_0)$ and that $\ell_n^P(\boldsymbol{\beta},\, \Lambda) = \ell_n^P(\boldsymbol{\beta}_0,\, \Lambda_0)$ with probability one under P_0, then we have $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\Lambda = \Lambda_0$, where ℓ_n^C and ℓ_n^P are the conditional and pairwise log-likelihood functions, respectively.*

Proof. For the detailed proof, please see the web supplementary materials. It is worth noting that $\Lambda_0(t)$ is not identifiable for $0 < t < w_1$ (Wang et al., 1993). However, since the support of A^* includes zero, by (C3), w_1 is usually close to zero; thus, the identifiability issue is less likely to occur. □

A.2 Consistency of $(\hat{\boldsymbol{\beta}}, \hat{\Lambda})$

To follow the consistency proof of general Z-estimators, we first bound the bracketing numbers (entropies) of the related bivariate function classes using the U-processes theory (De la Peña and Giné, 1999, Chapter 5). For $k = 0, 1, 2$, the classes $\{(\mathbf{z}_1, \mathbf{z}_2) \mapsto \mathbf{z}_1^{\otimes k} e^{\mathbf{z}_1^T \boldsymbol{\beta}} - \mathbf{z}_2^{\otimes k} e^{\mathbf{z}_2^T \boldsymbol{\beta}} : \mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^p;\ \boldsymbol{\beta} \in B\}$ are Euclidean (Nolan and Pollard, 1987); thus, their bracketing numbers in $L_1(P^2)$ are finite, where $P^2 \equiv P \otimes P$, and P is any probability measure on Ω. Bounds for bivariate function classes only consisting of indicator functions can be shown using the VC theory (see De la Peña and Giné, 1999, Section 5.2). Denoting the class of cumulative baseline hazard functions satisfying (C1) as $\mathcal{H}_\Lambda$, then we have

Lemma 2. *The bivariate function class* $\mathcal{H}_\Lambda^D = \{(s,t) \mapsto \Lambda(s) - \Lambda(t) : s, t \in [0, \tau];\ \Lambda \in \mathcal{H}_\Lambda\}$ *has finite bracketing numbers in* $L_1(P^2)$ *for all* $\varepsilon > 0$.

Proof. To avoid technicality, we assume all bivariate function classes involved in this and the following proofs are measurable (De la Peña and Giné, 1999, Section 3.5). Theorem 2.7.5 of van der Vaart and Wellner (1996) indicates that for any given $\varepsilon > 0$, there exists a constant K_1 such that the bracketing entropy $\log N_{[]}(\varepsilon, \mathcal{H}_\Lambda, L_1(P)) < K_1/\varepsilon < \infty$ for any probability measure P. For a given $\Lambda \in \mathcal{H}_\Lambda$, suppose an ε-bracket containing it in $L_1(P)$ is (Λ_l, Λ_u), we can show that $(\Lambda_l(s) - \Lambda_u(t),\ \Lambda_u(s) - \Lambda_l(t))$ is a 2ε-bracket for $\Lambda(s) - \Lambda(t)$ in $L_1(P^2)$. Since ε is arbitrary, $\mathcal{H}_\Lambda^D$ also has finite bracketing numbers in $L_1(P^2)$. $\square$

Remark 1. By Corollary 5.2.5 of De la Peña and Giné (1999), the finite bracketing numbers imply the corresponding function classes satisfy the uniform law of large numbers of U-processes. The uniform law of large numbers for $U^P(\boldsymbol{\beta}, \Lambda)$ and its derivatives then follow, because they are Lipschitz functions of the component functions with finite bracketing numbers (van der Vaart and Wellner, 1996).

Proof of Theorem 1. Since $\log \mathcal{L}_n^P$ is always negative, by the similar arguments as in Zeng and Lin (2006), we can show that the PLAC estimator has finite jump sizes, and that $\hat{\Lambda}(\tau)$ is bounded a.s. when $n \to \infty$. As in Section 2.3, we can write the composite score function as the summation of $U^C(\boldsymbol{\beta}, \Lambda)$ and $U^P(\boldsymbol{\beta}, \Lambda)$; the former is the conditional approach score function and has expectation zero at the true parameters. By double expectation, U^P also has mean zero at the true parameters. Let $N_i(s) = \Delta_i I(X_i \leqslant s)$ be the observed event counting process for subject i. Using the linear functional notations, the self-consistency solution of Λ is given by

$$\hat{\Lambda}(t) = \mathbb{P}_n \int_0^t \frac{dN(s)}{M_n(s; \hat{\boldsymbol{\beta}}, \hat{\Lambda})},$$

where

$$M_n(s; \hat{\boldsymbol{\beta}}, \hat{\Lambda}) = \mathbb{P}_n Y(s) e^{\mathbf{Z}^T \hat{\boldsymbol{\beta}}} + \mathbb{U}_{n,2} \frac{R(\hat{\boldsymbol{\beta}}, \hat{\Lambda})}{1 + R(\hat{\boldsymbol{\beta}}, \hat{\Lambda})} Q^{(0)}(s; \hat{\boldsymbol{\beta}}).$$

Inspired by the form of $\hat{\Lambda}(t)$, we define another random step function

$$\tilde{\Lambda}(t) = \mathbb{P}_n \int_0^t \frac{dN(s)}{M_n(s; \boldsymbol{\beta}_0, \Lambda_0)}.$$

Note that by (C2)-(C3), Lemma 2 and double expectation, the second term of $M_n(s; \boldsymbol{\beta}_0, \Lambda_0)$ is negligible compared with the first one when n is sufficiently large. Therefore, $M_n(s; \boldsymbol{\beta}_0, \Lambda_0)$ is finite and uniformly bounded away from 0 on $[0, \tau]$ as $n \to \infty$. Under the regularity conditions, by the Glivenko-Cantelli theorem, Remark 1 and the dominated convergence theorem, we can show $\left\| \tilde{\Lambda}(t) - \Lambda_0(t) \right\|_{L_\infty[0,\tau]} \to 0$ almost surely.

By the definition of the PLAC estimator, the log-composite-likelihood evaluated at $(\hat{\beta}, \hat{\Lambda})$ is greater than that evaluated at $(\beta_0, \tilde{\Lambda})$:

$$\begin{aligned} & \mathbb{P}_n \int_0^\tau \left\{ \log \frac{\hat{\Lambda}}{\tilde{\Lambda}}\{s\} + \mathbf{Z}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right\} dN(s) \\ -\mathbb{P}_n \left\{ e^{\mathbf{Z}^T \hat{\beta}} \int_0^\tau Y(s) d\hat{\Lambda}(s) - e^{\mathbf{Z}^T \beta_0} \int_0^\tau Y(s) d\tilde{\Lambda}(s) \right\} - \mathbb{U}_{n,2} \log \frac{1 + R(\hat{\boldsymbol{\beta}}, \hat{\Lambda})}{1 + R(\boldsymbol{\beta}_0, \tilde{\Lambda})} & \geqslant 0. \end{aligned}$$

Since $\boldsymbol{\beta}$ is in a compact set and that on $[0, \tau]$, and that $\hat{\Lambda}(t) \leqslant \hat{\Lambda}(\tau)$ is bounded with probability one, by the Helly's selection lemma, with probability one, for every subsequence of $(\hat{\boldsymbol{\beta}}, \hat{\Lambda})$, we can find a further subsequence along which $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}^*$ for some $\boldsymbol{\beta}^*$ and $\hat{\Lambda}(t) \to \Lambda^*(t)$, $\forall t \in [0, \tau]$ for some monotone function Λ^*.

By the absolute continuity of $\hat{\Lambda}(t)$ with respect to $\tilde{\Lambda}(t)$ and (C1), $\Lambda^*(t)$ is absolutely continuous with respect to the Lebesgue measure and we denote its derivative as $\lambda^*(t)$. Thus the ratio $d\hat{\Lambda}/d\tilde{\Lambda}$ converges to $\lambda^*(t)/\lambda_0(t)$. Again, by the Glivenko-Cantelli theorem, Remark 1 and the dominant convergence theorem, the difference of the log-compositelikelihoods converges to

$$P_0 \int_0^\tau \left\{ \log \frac{\lambda^*}{\lambda_0}(s) + \mathbf{Z}^T(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0) \right\} dN(s)$$
$$-P_0 \left\{ e^{\mathbf{Z}^T\boldsymbol{\beta}^*} \int_0^\tau Y(s) d\Lambda^*(s) - e^{\mathbf{Z}^T\boldsymbol{\beta}_0} \int_0^\tau Y(s) d\Lambda_0(s) \right\} - P_0 \log \frac{1 + R(\boldsymbol{\beta}^*, \Lambda^*)}{1 + R(\boldsymbol{\beta}_0, \Lambda_0)} \geqslant 0.$$

The left-hand side is the composite Kullback-Leibler divergence (Varin and Vidoni, 2005) of the density indexed by $(\boldsymbol{\beta}^*, \Lambda^*)$ from the true density, which by Lemma 1 should be strictly negative unless $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$ and $\Lambda^* = \Lambda_0$. Since every subsequence of $(\hat{\boldsymbol{\beta}}, \hat{\Lambda})$ has a further subsequence converging to $(\boldsymbol{\beta}_0, \Lambda_0)$, we have convergence of the entire sequence to the same limit. Finally, the uniform convergence of $\hat{\Lambda}(t)$ to $\Lambda_0(t)$ over $[0, \tau]$ follows from the continuity of Λ_0. □

A.3 Asymptotic Normality of $(\hat{\boldsymbol{\beta}}, \hat{\Lambda})$

We first establish a lemma on the $\sqrt{n}$-uniform convergence rate and asymptotic normality of the log-generalized odds ratio. This is achieved by the projection.

Lemma 3. *Under Conditions (C1)-(C4), the class of the log-generalized odds ratios*

$$\mathcal{R} = \{(\mathcal{O}_i, \mathcal{O}_j) \mapsto r_{ij}(\boldsymbol{\beta}, \Lambda) \equiv \log R_{ij}(\boldsymbol{\beta}, \Lambda) : \mathcal{O}_i, \mathcal{O}_j \in \Omega, \boldsymbol{\beta} \in B, \Lambda \in \mathcal{H}_\Lambda\}$$

satisfies $\sqrt{n}(\mathbb{U}_{n,2}r - P_0^2 r) \rightsquigarrow \mathbb{G}_r$, *where* $\mathbb{G}_r$ *is a tight mean-zero Gaussian process.*

Proof. We can show $\|\mathbb{U}_{n,2}r - P_0^2 r - \hat{\mathbb{U}}_{n,2}r\|_{\beta,\Lambda} = o_p(n^{-1/2})$, where $\hat{\mathbb{U}}_{n,2}r$ is the Hájek projection of $(\mathbb{U}_{n,2}r - P_0^2 r)$ (van der Vaart, 2000), and takes the form $\sum_{i=1}^n \mathrm{E}\left(\mathbb{U}_{n,2}r - P_0^2 r | \mathcal{O}_i\right)$. Since $\mathcal{O}_i$ and $\mathcal{O}_j$ are i.i.d., we have

$$\hat{\mathbb{U}}_{n,2}r = \frac{2}{n} \sum_{i=1}^n \left\{ e^{\mathbf{Z}_i^T\boldsymbol{\beta}} \Lambda(A_i) - \mathrm{E} e^{\mathbf{Z}_i^T\boldsymbol{\beta}} \cdot \Lambda(A_i) - e^{\mathbf{Z}_i^T\boldsymbol{\beta}} \mathrm{E}\Lambda(A_i) + \mathrm{E}(e^{\mathbf{Z}_i^T\boldsymbol{\beta}} \Lambda(A_i)) \right\} - 4P_0^2 r.$$

It can be verified that $P_0^2 r = 2\mathrm{Cov}(e^{\mathbf{Z}_i^T\boldsymbol{\beta}}, \Lambda(A_i))$; thus,

$$\tilde{\mathbb{U}}_{n,2} \equiv \mathbb{U}_{n,2}r - P^2 r - \hat{\mathbb{U}}_{n,2}r \asymp -2 \cdot \frac{1}{n} \sum_{i=1}^n (e^{\mathbf{Z}_i^T\boldsymbol{\beta}} - \mathrm{E}e^{\mathbf{Z}_i^T\boldsymbol{\beta}}) \cdot \frac{1}{n} \sum_{j=1}^n \left\{ \Lambda(A_j) - \mathrm{E}\Lambda(A_j) \right\},$$

where $\asymp$ means asymptotically equivalent. Note that both $\{\mathbf{z} \mapsto e^{\mathbf{z}^T\boldsymbol{\beta}} : \mathbf{z} \in \mathbb{R}^p, \boldsymbol{\beta} \in B\}$ and $\mathcal{H}_\Lambda$ are Donsker, thus we have

$$\left\|\tilde{\mathbb{U}}_{n,2}\right\|_{\boldsymbol{\beta},\Lambda} \lesssim \left\|n^{-1/2}\mathbb{G}_n e^{\mathbf{Z}^T\boldsymbol{\beta}}\right\|_{\boldsymbol{\beta}} \cdot \left\|n^{-1/2}\mathbb{G}_n\Lambda\right\|_{\Lambda} = O_p(n^{-1/2})O_p(n^{-1/2}) = o_p(n^{-1/2}),$$

where $\lesssim$ means the inequality holds up to a multiplicative constant and $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_0)$. Therefore, $\mathbb{U}_{n,2}r - P_0^2 r$ is equivalent to its projection $\hat{\mathbb{U}}_{n,2}r$ up to a $o_p(n^{-1/2})$ term. The weak convergence of $\hat{\mathbb{U}}_{n,2}r$ can be established using the empirical process theory. Combining these two facts leads to the weak convergence of $\mathbb{U}_{n,2}r$. □

Proof of Theorem 2. Let θ denote the parameters $(\boldsymbol{\beta}, \Lambda)$. We proceed by checking the four conditions in Theorem 3.3.1 of van der Vaart and Wellner (1996). Note that $\sqrt{n}U(\theta_0)$ can be decomposed into $\sqrt{n}U^C(\theta_0) + \sqrt{n}U^P(\theta_0)$. Following the martingale theory, the first term converges to a mean-zero Gaussian process $\mathbb{G}_{U^C}$, and the linear functional $\sqrt{n}\left\{b_1^T U_{\boldsymbol{\beta}}^C(\theta_0) + U_\Lambda^C(\theta_0)(h)\right\}$ converges to a mean-zero normal random variable with the variance that can be consistently estimated by $b^T\hat{V}^C b$, where b is defined as in Section 2.3. For the second term, by Lemma 3, the preservation theorem of Lipschitz functions and Theorem 5.3.1 of De la Peña and Giné (1999), it also converges to a mean-zero Gaussian process $\mathbb{G}_{U^P}$, and $\sqrt{n}\left\{b_1^T U_{\boldsymbol{\beta}}^P(\theta_0) + U_\Lambda^P(\theta_0)(h)\right\}$ converges to a mean-zero normal random variable with the variance that can be consistently estimated by $b^T\hat{V}^P b$. Given $\{(A_i, \mathbf{Z}_i)\}_{i=1}^n$, $U^C(\theta_0)$ is a martingale, whereas $U^P(\theta_0)$ is a function of A_i and $\mathbf{Z}_i$ only; thus, by double expectation,

$$\mathrm{E}_0\left(U^C(\theta_0)\cdot U^P(\theta_0)\right) = \mathrm{E}_0\left\{\mathrm{E}_0\left(U^C(\theta_0)|\{(A_i,\mathbf{Z}_i)\}_{i=1}^n\right)\cdot U^P(\theta_0)\right\} = \mathrm{E}_0\left(0\cdot U^P(\theta_0)\right) = 0,$$

where $\cdot$ denotes the inner product. This indicates that $U^C(\theta_0)$ and $U^P(\theta_0)$ are asymptotically independent (van der Vaart and Wellner, 1996, Example 1.4.6) at θ_0 and that $\sqrt{n}U(\theta_0)$ converges to a mean-zero Gaussian process $\mathbb{G}_U$. In addition, the linear functional $\sqrt{n}\left\{b_1^T U_{\boldsymbol{\beta}}(\theta_0) + U_\Lambda(\theta_0)(h)\right\}$ converges to a mean-zero normal random variable with asymptotic variance that can be consistently estimated by $b^T(\hat{V}^C + \hat{V}^P)\,b$. Therefore, the two stochastic conditions are satisfied by the consistency of $\hat{\theta}$, Lemma 3 and Lemma 3.3.5 of van der Vaart and Wellner (1996). The fourth condition holds since $\hat{\theta}$ is a zero of $U(\theta)$, and that $u(\theta_0) \equiv \mathrm{E}_0 U(\theta_0) = 0$ by the arguments in the consistency proof.

The Fréchet-differentiability can be checked directly. For the continuous invertibility, similar to $U(\theta_0)$, we decompose $J \equiv \partial u(\theta)/\partial\theta|_{\theta=\theta_0}$ into J^C and J^P. By (C5) and the classic Cox model results, J^C is continuously invertible. Thus, it suffices to show J^P is a compact operator and that J is one-to-one by the Fredholm theory. The former can be shown by the Helly's lemma and the compactness of the bounded linear operator with finite dimensional range. To show J is one-to-one, we follow the similar arguments as in Zeng and Lin (2006). For $(b, h) \in \mathbb{R}^p \times BV[0, \tau]$, we need to show $J(b, h) = 0$ implies $b = 0$ and $h(t) = 0$, where

$$\begin{aligned} J(b,h) &= P_0\left\{\left(b^T\int \mathbf{Z}(dN - Ye^{\mathbf{Z}^T\beta_0}d\Lambda_0) + \int hdN - \int Ye^{\mathbf{Z}^T\beta_0}hd\Lambda_0\right)^2 \right.\\ &\quad \left. + \frac{1}{R_0}\left(\frac{R_0}{1+R_0}b^T\int Q_0^{(1)}d\Lambda_0 + \frac{R_0}{1+R_0}\int Q_0^{(0)}hd\Lambda_0\right)^2\right\}. \end{aligned}$$

From the preceding display, we find that $J(b, h) = 0$ indicates the conditional score along the path $(\boldsymbol{\beta}_0 + b, \Lambda_0 + \varepsilon\int hd\Lambda_0)$ is zero with probability one, i.e.,

$$b^T\int_0^\tau \mathbf{Z}[dN(s) - Y(s)e^{\mathbf{Z}^T\beta_0}d\Lambda_0(s)] + \int_0^\tau h(s)dN(s) - \int_0^\tau Y(s)e^{\mathbf{Z}^T\beta_0}h(s)d\Lambda_0(s) = 0.$$

By (C1)-(C3), considering the case of $N(\tau) = 0$ and $A + C \geqslant \tau$ and the case of $N(t) = I(t \geqslant t_0)$, $t_0 \in [0, \tau]$ and $A + C \geqslant \tau$, we have $b = 0$ and $h(t) = 0$.

With all four conditions satisfied, by Theorem 3.3.1 of van der Vaart and Wellner (1996), we have $\sqrt{n}(\hat\theta - \theta_0) \rightsquigarrow -J^{-1}\mathbb{G}_U$, Since linear maps preserve the Gaussian property, $\sqrt{n}(\hat\theta - \theta_0)$ also converge weakly to a mean-zero Gaussian process, and the asymptotic variance of the linear functional (8) is given by (9). The consistency of (9) can be clearly shown by the Glivenkon-Cantelli theorems and Remark 1. $\square$

References

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, pages 1100–1120.

Asgharian, M., Wolfson, D. B., and Zhang, X. (2006). Checking stationarity of the incidence rate using prevalent cohort survival data. *Statistics in Medicine*, 25(10):1751–1767.

Borgan, Ø. and Liestøl, K. (1990). A note on confidence intervals and bands for the survival function based on transformations. *Scandinavian Journal of Statistics*, pages 35–41.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 34(2):187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

De la Peña, V. and Giné, E. (1999). *Decoupling: from dependence to independence*. Springer.

Huang, C.-Y. and Qin, J. (2012). Composite partial likelihood estimation under length-biased sampling, with application to a prevalent cohort study of dementia. *Journal of the American Statistical Association*, 107(499):946–957.

Huang, C.-Y. and Qin, J. (2013). Semiparametric estimation for the additive hazards model with left-truncated and right-censored data. *Biometrika*, 100(4):877–888.

Huang, C.-Y., Qin, J., and Follmann, D. A. (2012). A maximum pseudo-profile likelihood estimator for the cox model under length-biased sampling. *Biometrika*, 99(1):199–210.

Jung, S.-H. (1999). Rank tests for matched survival data. *Lifetime Data Analysis*, 5(1):67–79.

Kalbfleisch, J. and Lawless, J. (1991). Regression models for right truncated data with applications to aids incubation times and reporting lags. *Statistica Sinica*, 1:19–32.

Kalbfleisch, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 32(2):175–208.

Liang, K.-Y. and Qin, J. (2000). Regression analysis under non-standard situations: a pairwise pseudolikelihood approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):773–786.

Mandel, M. and Betensky, R. A. (2007). Testing goodness of fit of a uniform truncation model. *Biometrics*, 63(2):405–412.

Murphy, S., Rossini, A., and van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, 92(439):968–976.

Ning, J., Qin, J., and Shen, Y. (2011). Buckley–James-type estimator with right-censored and length-biased data. *Biometrics*, 67(4):1369–1378.

Ning, J., Qin, J., and Shen, Y. (2014). Semiparametric accelerated failure time model for length-biased data with application to dementia study. *Statistica Sinica*, 24(1):313.

Nolan, D. and Pollard, D. (1987). U-processes: rates of convergence. *The Annals of Statistics*, pages 780–799.

Perlman, R. L., Kiser, M., Finkelstein, F., Eisele, G., Roys, E., Liu, L., Burrows-Hudson, S., Port, F., Messana, J. M., Bailie, G., et al. (2003). Renal research institute symposium: The longitudinal chronic kidney disease study: A prospective cohort study of predialysis renal failure. *Seminars in Dialysis*, 16(6):418–423.

Pollard, D. (2002). *A user's guide to measure theoretic probability*, volume 8. Cambridge University Press.

Qin, J. and Liang, K. (1999). Generalized odds ratio model and pairwise conditional likelihood. Technical report, Technical Report. Department of Biostatistics, Johns Hopkins University, Baltimore, MD.

Qin, J., Ning, J., Liu, H., and Shen, Y. (2011). Maximum likelihood estimations and em algorithms with length-biased data. *Journal of the American Statistical Association*, 106(496):1434–1449.

Qin, J. and Shen, Y. (2010). Statistical methods for analyzing right-censored length-biased data under cox model. *Biometrics*, 66(2):382–392.

Sen, P. K. (1960). On some convergence properties of u-statistics. *Calcutta Statist. Assoc. Bull*, 10(1):18.

Shen, Y., Ning, J., and Qin, J. (2009). Analyzing length-biased data with semiparametric transformation and accelerated failure time models. *Journal of the American Statistical Association*, 104(487):1192–1202.

Tsai, W. Y. (2009). Pseudo-partial likelihood for proportional hazards models with biased-sampling data. *Biometrika*, 96(3):601–615.

Tsai, W.-Y., Jewell, N. P., and Wang, M.-C. (1987). A note on the product-limit estimator under right censoring and left truncation. *Biometrika*, 74(4):883–886.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 38(3):290–295.

van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence*. Springer.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *The Annals of Statistics*, 10(2):616–620.

Vardi, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: nonparametric estimation. *Biometrika*, 76(4):751–761.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42.

Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika*, 92(3):519–528.

Wang, M.-C. (1996). Hazards regression analysis for length-biased data. *Biometrika*, 83(2):343–354.

Wang, M.-C., Brookmeyer, R., and Jewell, N. P. (1993). Statistical models for prevalent cohort data. *Biometrics*, 49(1):1–11.

Wang, M.-C., Jewell, N. P., and Tsai, W.-Y. (1986). Asymptotic properties of the product limit estimate under random truncation. *The Annals of Statistics*, pages 1597–1605.

Zeng, D. and Lin, D. (2006). Efficient estimation of semiparametric transformation models for counting processes. *Biometrika*, 93(3):627–640.